SUBSYSTEM ENCODING AND CROSS-GENOME PROJECTION USING THE SEED
Tutorial. University of Florida at Gainesville, March 02-04, 2005

INTRODUCTION

Our goals for this tutorial are:

- **to introduce a subsystems approach** to genome comparative analysis and annotation, a key technology aspect of the *Project to Annotate the first 1,000 Genomes* launched by the Fellowship for Interpretation of Genomes (FIG, Bur Ridge, IL);
- **to provide basic training** in using the SEED genomic data base and tools developed by FIG;
- **to help you to encode, explore and extend a subsystem** of your choice (from the list provided).

Our experience suggests that working on such an assignment is the most efficient way to learn the tool and to grasp the power of the approach, which, we believe, will be very useful in your research work. Accomplishing an assignment will allow you to reliably predict how your subsystem is implemented in those organisms where almost no experimental data is available. You will significantly refine genomic annotations, reconcile inconsistencies and reveal open problems. Quite likely, you will find several cases of "missing genes" or unclear biochemistry. With persistence and a certain amount of luck you will come up with predictions that can be further tested experimentally. By "adopting" a subsystem for further curation you will join the ranks of scientists who are using SEED tools to capture and share their expertise and to support their own research. This effort may become a starting point for a review article or a research project.

In preparation for the tutorial we encourage you:

- **To browse through a draft of our Subsystems Paper.** Follow the link in the UofFloridaTutorial Main Page to download the document. It will provide you with a brief overview of the Project, with introduction to concepts and terms (e.g. "subsystem") and with an illustration of developed subsystems. It will help you to be better prepared to the challenges of the upcoming tutorial.
- **To choose and claim a subsystem** from the Table on the UofFloridaTutorial Main Page Main Page. Follow the guidelines described in the assignment below. Highly recommended for all participants, mandatory for students.
- **To study and understand the biology and biochemistry of the chosen subsystem** before the tutorial. You do not want to spend any time on the basics of your subsystem during the tutorial, where you will be busy enough learning how to use tools, exploring convoluted cases, resolving annotation conflicts etc. By "doing your homework" at home, you will maximize the efficiency of the tutorial, as well as increase the odds of getting good grades (when applicable).
- **Capture your understanding of a subsystem** in a form described below. For a typical set of subsystem preparation materials, follow the link on the UofFloridaTutorial Main Page. Feel free to use these files as templates.

ASSIGNMENT

Your work on a subsystem will proceed in THREE STEPS (two of them are mandatory for students).

STEP 1. BEFORE THE TUTORIAL (mandatory);
STEP2. DURING THE TUTORIAL (mandatory);
STEP3. AFTER THE TUTORIAL (optional)

Here we will focus mostly on STEP1. The other two will be fully covered in the class.

STEP 1. BEFORE THE TUTORIAL

- **choose and claim a subsystem for encoding.**
    - Examine the table where subsystems are listed along with some features.
    - For those that appear attractive (or even neutral), follow the links to another web-site (SUBSYSTEM FORUM).
    - A first posting in each thread provides you with a very brief introduction of a subsystem and a scope of work. The most important of all are "starting point" references. Find the respective papers, and use them to assess a task. Some threads may contain additional postings. They are there to help you with the subsystem encoding once it is finally assigned.
    - Settle with the first three choices and e.mail them to Andrei (osterman@burnham.org) ASAP.
    - We will assign the subsystems on the "first come-first serve" basis.
    - A failure to send us your choices within the first 24 hrs (from the first e.mail announcement), will be interpreted as "no preferences" and followed by an arbitrary assignment of the residual subsystems.
    - In any event you may be able to renegotiate the subsystem based on availability.
    - You have an option to suggest an alternative subsystem (use the same format as in postings, include a reference and comments). Your suggestion will be considered and, most likely, accepted except for the cases of obvious redundancy. Feel free to consult with Subsystem BulletinBoard for the list of existing and developing subsystems (http://www-unix.mcs.anl.gov/SEEDWiki/moin.cgi/SubsystemBulletinBoard).
    - Important: subsystems inevitably differ in complexity (topology, gaps in knowledge) and size (a number of reactions and enzymes). Students will be graded based on the effort and insight rather than on completion. Therefore a partial coverage of a complex subsystem may be regarded as highly as complete coverage of a simpler one.

- **Prepare supporting materials for subsystem encoding.**
    - Once a subsystem is assigned to you (your name will appear in the column "Claimed by" of the Table), start working on supporting materials.
    - Important: this is one of the most critical aspects of the assignment. More generally, with adequate tools (such as SEED), encoding and extension of a subsystem is largely a technical task with an almost guaranteed success. These tools are aimed to capture and enhance, not to replace, preexisting knowledge and understanding. The more you know about a basic version of a subsystem before the tutorial, the better.
    - You do not need to use SEED or any other specialized databases to prepare for the assignment, although feel free to do so if you wish. Original papers, review articles, especially those helping to define pathways, reactions, individual functional roles and examples of genes from model species is all that you need. Start by recommended references and extend the search using references therein, PubMed and so forth.
    - Among web-resources, KEGG and MetaCyc may (or may not) be helpful but ONLY AS A SECONDARY SOURCE. Our goal is to go from original findings, not from their encoding in other systems, which in some cases may be dangerous (propagation of errors, etc). Books listed in the end may be helpful in some cases.
    - A complete set of supporting materials should include
        - **"Free text" notes** (use any word processor) about the subsystem, most important fundamental and applied aspects, occurrence in various species, open problems. Carefully selected and annotated reference list. These notes will be expanded as you will encode and explore a subsystem. They will be further incorporated in SEED and SUBSYSTEM FORUM.

    - **List of functional roles, gene examples.** May be included with text notes, or separately, as Excel file or alike. Names of roles will not be final, and you should not spend too much time thinking about them yet. We will address SEED naming conventions and rules (to the extent they even exist) at the tutorial. Record any gene IDs as provided in papers. Genbank, Swissprot or FASTA sequences will work equally well.
    - **Sketch of a subsystem diagram.** All that matters is your understanding of connectivity in a subsystem, sequence of reactions, intermediates, interactions between protein components, alternative routes and isozymes. Feel free to use a powerpoint diagram (in the sample folder) as a template. Otherwise use any graphic environment you are comfortable with. Screengrabs, PDF fragments, JPEG figures from papers (as in your assignments) will be enough to get started. Everything else can be added later.
    - **Relevant articles.** It is advisable to have hard copies (and/or PDF files) of several most relevant publications to use them in the class.
- You will benefit from having most of these materials in electronic format. Bring them on a disc or a USB-memory stick, send them by e.mail to us and/or to yourself, consult with IT people at the University on how to set up a folder that you would be able to access from terminals in the class.

This is all you need to know and do before the tutorial. Following these guidelines is a key to your most successful performance. Ignoring them is the most certain path to failure.

Questions about the assignment? Contact Andrei Osterman (osterman@burnham.org).
Questions about logistics, access to local instances of SEED? Contact Dr. Valerie de Crecy Lagard at U of F (vcrecy@ufl.edu). Meanwhile, you may always browse an open remote instance at http://TheSEED.uchicago.edu/FIG/index.cgi.

STEP2. DURING THE TUTORIAL

- The tutorial will start with a brief introduction of the Project, approach, tools and applications. Our goal will be to move very quickly to a hands-on subsystem encoding. You will learn the tools by using them, by making mistakes and seeking help, as opposed to listening and watching.
- During the whole tutorial, our team will provide you with all the technical help you may need. However, with respect to specific biology of your subsystem, you are on your own.
- Within the time provided, your task will be to encode a subsystem using SEED tools. You will start by enlisting functional roles and connecting them to genes in 1 or 2 model species covered by available references.
- You will continue by expanding a subsystem to a series of 10-20 related and remote complete genomes. Each addition will include annotating genes using the whole variety of specialized tools in SEED. You will assess functional and non-functional variants as implemented in various species based on their gene content.
- Further expansion of a subsystem will reveal characteristic patterns as well as inconsistencies and deficiencies. Some of them will be easy to reconcile. Others will constitute genuine gaps of knowledge, such as missing genes.
- We will introduce you to *genome context analysis* tools, which will help you to improve annotations and, in some cases, will allow you to reveal candidate genes to fill-in gaps (for a review of the approach and examples, see Osterman, A. & Overbeek R. (2003) “Missing genes in metabolic pathways: A comparative genomics approach.” *Current Opin. Chem. Biol.* **7,** 1-14.)

STEP3. AFTER THE TUTORIAL

- The results of subsystem development by students, which include a subsystem encoding in SEED environment and web-posting of supporting materials (such as diagrams and notes), will be assessed and graded.
- We believe that some of you will get interested and proud enough of your accomplishments to continue a subsystem development and curation beyond this tutorial. We will discuss all the technical and conceptual details of this opportunity.

We are looking forward to having the most productive and exciting workshop!

Recommended books for the first step of subsystem analysis:

1. Gerhard Michal. "Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology"
2. "Escherichia Coli and Salmonella: Typhimurium Cellular and Molecular Biology", Ed. Neidhard (also available on CD)
3. "Bacillus Subtilis and Its Closest Relatives: From Genes to Cells" by Abraham L. Sonenshein et al
4. David White "The Physiology and Biochemistry of Prokaryotes";
5. Gerhard Gottschalk "Bacterial Metabolism"
6. Search in "PROKARYOTES": at http://141.150.157.117:8080/prokPUB/index.htm